Resampling Fewer Than n Observations: Gains, Losses, and Remedies for Losses

By

P.J. Bickel[(1),(2),(3)]

F. Götze[(1)]

W.R. van Zwet[(2)]

Research supported by
(1) NATO Grant CRG 920650
(2) Sonderforschungsbereich 343 Diskrete Strukturen der Mathematik, Bielefeld
(3) NSA Grant MDA 904-94-H-2020

Abstract: We discuss a number of resampling schemes in which $m = o(n)$ observations are resampled. We review nonparametric bootstrap failure and give results old and new on how the m out of n with replacement bootstraps and without replacement works. We extend work of Bickel and Yahav (1988) to show that m out of n bootstraps can be made second order correct, if the usual nonparametric bootstrap is correct and study how these extrapolation techniques work when the nonparametric bootstrap doesn't.

Key words and phrases: Asymptotic, bootstrap, nonparametric, parametric, testing.

1. Introduction

Over the last 10-15 years Efron's nonparametric bootstrap has become a general tool for setting confidence regions, prediction, estimating misclassification probabilities, and other standard exercises of inference when the methodology is complex. Its theoretical justification is based largely on asymptotic arguments for its consistency or optimality. A number of examples have been addressed over the years in which the bootstrap fails asymptotically. Practical anecdotal experience seems to support theory in the sense that the bootstrap generally gives reasonable answers but can bomb.

In a recent paper Politis and Romano (1992), following Wu (1990), and independently Götze (1993) showed that what we call the m out of n without replacement bootstrap with $m = o(n)$ typically works to first order both in the situations where the bootstrap works and where it does not.

The m out of n with replacement bootstrap with $m = o(n)$ has been known to work in all known realistic examples of bootstrap failure. In this paper,

- We show the large extent to which the Politis, Romano, Götze property is shared by the m out of n with replacement bootstrap and show that the latter has advantages.
- If the usual bootstrap works the m out of n bootstraps pay a price in efficiency. We show how, by the use of extrapolation the price can be avoided.
- We support some of our theory with simulations.

The structure of our paper is as follows. In section 2 we review a series of examples of success and failure to first order (consistency) of (Efron's) nonparametric bootstrap (nonparametric). We try to isolate at least heuristically some causes of nonparametric bootstrap failure. Our framework here is somewhat novel. In Section 3 we formally introduce the m out of n with and without replacement bootstrap as well as what we call "sample splitting", and establish their first order properties restating the Politis-Romano-Götze result. We relate these approaches to smoothing methods. Section 4 establishes the deficiency of the m out of n bootstrap to higher order if the nonparametric bootstrap works to first order and section 5 shows how to remedy this deficiency to second order by extrapolation. In section 6 we study how the improvements of section 5 behave when the nonparametric bootstrap doesn't work to first order. We present simulations in section 7 and proofs of our new results in section 8. The critical issue of choice of m and applications to testing will be addressed elsewhere.

2. Successes and failure of the bootstrap

We will limit our work to the iid case because the issues we discuss are clearest in this context. Extension to the stationary mixing case, as done for the m out of n without replacement bootstrap in Politis and Romano (1994), are possible but the study of higher order properties as in sections 4 and 5 of our paper is more complicated.

We suppose throughout that we observe $X_1, \ldots, X_n$ taking values in $\boldsymbol{X} = R^p$ (or more generally a separable metric space.) i.i.d. according to $F \in \mathcal{F}_0$. We stress that $\mathcal{F}_0$ need not be and usually isn't the set of all possible distributions. In hypothesis testing applications, $\mathcal{F}_0$ is the hypothesized set, in looking at the distributions of extremes, $\mathcal{F}_0$ is the set of populations for which extremes have limiting distributions. We are interested in the distribution of a symmetric function of $X_1, \ldots, X_n$; $T_n(X_1, \ldots, X_n, F) \equiv T_n(\hat{F}_n, F)$ where $\hat{F}_n$ is defined to be the empirical distribution of the data. More specifically we wish to estimate a parameter which we denote $\theta_n(F)$, of the distribution of $T_n(\hat{F}_n, F)$, which we denote by $\mathcal{L}_n(F)$. We will usually think of θ_n as real valued, for instance, the variance of $\sqrt{n}$ median $(X_1, \ldots, X_n)$ or the 95% quantile of the distribution of $\sqrt{n}(\bar{X} - E_F(X_1))$.

Suppose $T_n(\cdot, F)$ and hence θ_n is defined naturally not just on $\mathcal{F}_0$ but on $\mathcal{F}$ which is large enough to contain all discrete distributions. It is then natural to estimate F by the nonparametric maximum likelihood estimate, (NPMLE), $\hat{F}_n$, and hence $\theta_n(F)$ by the plug in $\theta_n(\hat{F}_n)$. This is Efron's (ideal) nonparametric bootstrap. Since $\theta_n(F) \equiv \gamma(\mathcal{L}_n(F))$ and, in the cases we consider, computation of γ is straightforward the real issue is estimation of $\mathcal{L}_n(F)$. Efron's (ideal) bootstrap is to estimate $\mathcal{L}_n(F)$ by the distribution of $T_n(X_1^*, \ldots, X_n^*, \hat{F}_n)$ where, given $X_1, \ldots, X_n$, the X_i^* are i.i.d. $\hat{F}_n$, i.e. the bootstrap distribution of T_n. In practice, the bootstrap distribution is itself estimated by Monte Carlo or more sophisticated resampling schemes, (see deCiccio and Romano (1988) and Hinkley (1988)). We will not enter into this question further.

Theoretical analyses of the bootstrap and its properties necessarily rely on asymptotic theory, as $n \to \infty$ coupled with simulations. We restrict analysis to $T_n(\hat{F}_n, F)$ which are asymptotically stable and nondegenerate on $\mathcal{F}_0$. That is, for all $F \in \mathcal{F}_0$, at least weakly

$$\mathcal{L}_n(F) \to \mathcal{L}(F) \text{ non degenerate}$$

$$\theta_n(F) \to \theta(F) \tag{2.1}$$

as $n \to \infty$.

Using m out of n bootstraps or sample splitting implicitly changes our goal from estimating features of $\mathcal{L}_n(F)$ to features of $\mathcal{L}_m(F)$. This is obviously nonsensical without assuming that the laws converge.

Requiring non degeneracy of the limit law means that we have stabilized the scale of $T_n(\hat{F}_n, F)$. Any functional of $\mathcal{L}_n(F)$ is also a functional of the distribution of $\sigma_n T_n(\hat{F}_n, F)$ where $\sigma_n \to 0$ which also converges in law to point mass at 0. Yet this degenerate limit has no functional $\theta(F)$ of interest.

Finally, requiring that stability need occur only on $\mathcal{F}_0$ is also critical since failure to converge off $\mathcal{F}_0$ in a reasonable way is the first indicator of potential bootstrap failure.

2.1. When does the nonparametric bootstrap fail?

If θ_n doesn't depend on n, the bootstrap works, (is consistent on $\mathcal{F}_0$), if θ is continuous at all points of $\mathcal{F}_0$ with respect to weak convergence on $\mathcal{F}$. Conversely, the nonparametric bootstrap can fail if,

1. θ is not continuous on $\mathcal{F}_0$.

   An example we explore later is $\theta_n(F) = 1(F \text{ discrete})$ for which $\theta_n(\hat{F}_n)$ obviously fails if F is continuous.

   Dependence on n introduces new phenomena. In particular, here are two other reasons for failure we explore below.

2. θ_n is well defined on all of $\mathcal{F}$ but θ is defined on $\mathcal{F}_0$ only or exhibits wild discontinuities when viewed as a function on $\mathcal{F}$. This is the main point of examples 3-6.

3. $T_n(\hat{F}_n, F)$ is not expressible as or approximable on $\mathcal{F}_0$ by a continuous function of $\sqrt{n}(\hat{F}_n - F)$ viewed as an object weakly converging to a Gaussian limit in a suitable function space. See Giné and Zinn (1989). Example 7 illustrates this failure. Again this condition is a diagnostic and not necessary for failure as example 6 shows.

We illustrate our framework and discuss prototypical examples of bootstrap success and failure.

2.2. Examples of bootstrap success

Example 1 Confidence intervals: Suppose $\sigma^2(F) \equiv \mathrm{Var}_F(X_1) < \infty$ for all $F \in \mathcal{F}_0$.

a) Let $T_n(\hat{F}_n, F) \equiv \sqrt{n}(\bar{X} - E_F X_1)$. For the percentile bootstrap we are interested in $\theta_n(F) \equiv P_F[T_n(\hat{F}_n, F) \leq t]$. Evidently $\theta(F) = \Phi\left(\frac{t}{\sigma(F)}\right)$. In fact, we want to estimate the quantiles of the distribution of $T_n(\hat{F}_n, F)$. If $\theta_n(F)$ is the $1-\alpha$ quantile then $\theta(F) = \sigma(F) z_{1-\alpha}$ where z is the Gaussian quantile
b) Let $T_n(\hat{F}_n, F) = \sqrt{n}(\bar{X} - E_F X_1)/s$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. If $\theta_n(F) \equiv P_F(T_n(\hat{F}_n, F) \leq t]$ then, $\theta(F) = \Phi(t)$, independent of F. It seems silly to be estimating a parameter whose value is known but of course, interest now centers on $\theta'(F)$ the next higher order term in $\theta_n(F) = \Phi(t) + \frac{\theta'(F)}{\sqrt{n}} + O(n^{-1})$. □

Example 2 Estimation of variances: Suppose F has unique median $m(F)$, continuous density $f(m(F)) > 0$, $E_F|X|^\delta < \infty$, some $\delta > 0$ for all $F \in \mathcal{F}_0$ and $\theta_n(F) = \mathrm{Var}_F(\sqrt{n}$

median $(X_1, \ldots, X_n)$). Then $\theta(F) = [4f^2(m(F))]^{-1}$ on $\mathcal{F}_0$.

Note that, whereas θ_n is defined for all empirical distributions F in both examples the limit $\theta(F)$ is 0 or ∞ for such distributions in the second. Nevertheless, it is well known, see Efron (1979), that the nonparametric bootstrap is consistent in both examples in the sense that $\theta_n(\hat{F}_n) \stackrel{P}{\to} \theta(F)$ for $F \in \mathcal{F}_0$. □

2.3. Examples of bootstrap failure

Example 3 Confidence bounds for an extremum: This is a variation on Bickel-Freedman (1981). Suppose that all $F \in \mathcal{F}_0$ have a density f continuous and positive at $F^{-1}(0) > -\infty$. It is natural to base confidence bounds for $F^{-1}(0)$ on the bootstrap distribution of

$$T_n(\hat{F}_n, F) = n(\min_i X_i - F^{-1}(0)).$$

Let

$$\theta_n(F) = P_F[T_n(\hat{F}_n, F) > t] = (1 - F(\frac{t}{n} + F^{-1}(0))^n.$$

Evidently

$$\theta_n(F) \to \theta(F) = \exp(-f(F^{-1}(0))t)$$

on $\mathcal{F}_0$.

The nonparametric bootstrap fails. Let

$$N_n^*(t) = \sum_{i=1}^{n} 1(X_i^* \leq \frac{t}{n} + X_{(1)}), t > 0$$

where $X_{(1)} \equiv \min_i X_i$. Given $X_{(1)}$, $n\hat{F}_n(\frac{t}{n} + X_{(1)})$ is distributed as 1+ binomial $(n-1, \frac{F(\frac{t}{n}+X_{(1)})-F(X_{(1)})}{(1-F(X_{(1)}))})$ which converges weakly to a Poisson $(f(F^{-1}(0))t)$ variable. More generally, $n\hat{F}_n(\frac{\cdot}{n} + X_{(1)})$ converges weakly conditionally to $1 + N(\cdot)$, where N is a homogeneous Poisson process with parameter $f(F^{-1}(0))$. It follows that $N_n^*(\cdot)$ converges weakly (marginally) to a process $M(1 + N(\cdot))$ where M is a standard Poisson process independent of $N(\cdot)$. Thus if, in Efron's notation, we use P^* to denote conditional probability given $\hat{F}_n$ and let $\hat{F}_n^*$, be the empirical d.f. of $X_1^*, \ldots, X_n^*$ then

$$P^*[T_n(\hat{F}_n^*) > t] = P^*[N_n^*(t) = 0]$$

converges weakly to the random variable $P[M(1+N(t))=0|N]=e^{-(N(t)+1)}$ rather than to the desired $\theta(F)$. $\square$

Example 4 Extrema for unbounded distributions: (Athreya and Fukuchi (1994), Deheuvels, Mason, Shorack (1993))

Suppose $F \in \mathcal{F}_0$ are in the domain of attraction of an extreme value distribution. That is: for some constants $A_n(F)$, $B_n(F)$,

$$n(1-F)(A_n(F)+B_n(F)x) \to H(x,F)$$

where H is necessarily one of the classical three types (David (1981), p.259). $e^{-\beta x}1(\beta x \geq 0)$, $\alpha x^{-\beta}1(x \geq 0)$, $\alpha(-x)^{\beta}1(x \leq 0)$, for $\alpha, \beta \neq 0$. Let,

$$\theta_n(F) \equiv P[(\max(X_1,\ldots,X_n) - A_n(F))/B_n(F) \leq t] \to e^{-H(t,F)} \equiv \theta(F). \tag{2.2}$$

Particular choices of $A_n(F)$, for example, $F^{-1}(1-\frac{1}{n})$ and $B_n(F)$ are of interest in inference. However, the bootstrap doesn't It is easy to see that

$$n(1-\hat{F}_n(A_n(F)+tB_n(F))) \xrightarrow{w} N(t) \tag{2.3}$$

where N is an inhomogeneous Poisson process with parameter $H(t,F)$. Hence if $T_n(\hat{F}_n, F) = (\max(X_1,\ldots,X_n) - A_n(F))/B_n(F)$ then

$$P^*[T_n(\hat{F}_n^*, F) \leq t] \stackrel{w}{\Rightarrow} e^{-N(t)}. \tag{2.4}$$

It follows that the nonparametric bootstrap is inconsistent for this choice of A_n, B_n. If it were consistent, then

$$P^*[T_n(\hat{F}_n^*, \hat{F}_n) \leq t] \xrightarrow{P} e^{-H(t,F)} \tag{2.5}$$

for all t and (2.5) would imply that it is possible to find random A real and $B \neq 0$ such that

$$N(Bt+A) = H(t,F)$$

with probability 1. But $H(t,F)$ is continuous except at 1 point. So (2.4) and (2.5) contradict each other. Again, $\theta(F)$ is well defined for $F \in \mathcal{F}_0$ but not otherwise. Furthermore, small perturbations in F can lead to drastic changes in the nature of H, so that θ is not continuous if $\mathcal{F}_0$ is as large as possible.

Essentially the same bootstrap failure arises when we consider estimating the mean of distributions in the domain of attraction of stable laws of index $1 < \alpha \leq 2$. See Athreya (1987).

Example 5 Testing and improperly centered U and V statistics: (Bretagnolle (1983))

Let $\mathcal{F}_0 = \{F : F[-c, c] = 1, E_F X_1 = 0\}$ and let $T_n(\hat{F}_n) = n\bar{X}^2 = n \int xy d\hat{F}_n(x) d\hat{F}_n(y)$. This is a natural test statistic for $H : F \in \mathcal{F}_0$. Can one use the nonparametric bootstrap to find the critical value for this test statistic? Intuitively, $\hat{F}_n \notin \mathcal{F}_0$ and this procedure is rightly suspect. Nevertheless, in more complicated contexts, it is a mistake made in practice. David Freedman pointed us to Freedman et al (1994) where the Bureau of the Census appears to have fallen into such a trap – see Hall and Wilson (1991) for other examples. The nonparametric bootstrap may, in general, not be used for testing as will be shown in a forthcoming paper.

In this example, due to Bretagnolle (1983), we focus on $\mathcal{F}_0$ for which a general U or V statistic T is degenerate and show that the nonparametric bootstrap doesn't work. More generally, suppose $\psi : R^2 \to R$ is bounded and symmetric and let $\mathcal{F}_0 = \{F : \int \psi(x, y) dF(x) = 0 \text{ for all } y\}$.

Then, it is easy to see that

$$T_n(\hat{F}_n) = \int \psi(x, y) dW_n^0(x) dW_n^0(y) \tag{2.6}$$

where $W_n^0(x) \equiv \sqrt{n}(\hat{F}_n(x) - F(x))$ and well known that

$$\theta_n(F) \equiv P_F[T_n(\hat{F}_n) \leq t] \to P[\int \psi(xy) dW^0(F(x)) dW^0(F(y)) \leq t] \equiv \theta(F)$$

where W^0 is a Brownian Bridge. On the other hand it is clear that,

$$\begin{aligned} T_n(\hat{F}_n^*) &= n \int \psi(x, y) d\hat{F}_n^*(x) d\hat{F}_n(y) \\ &= \int \psi(x, y) dW_n^*(x) dW_n^{0*}(y) \end{aligned} \tag{2.7}$$

$$+2\int \psi(x,y)dW_n^0(x)dW_n^{0*}(y)$$
$$+\int \psi(x,y)dW_n^0(x)dW_n^0(y)$$

where $W_n^{0*}(x) \equiv \sqrt{n}(\hat{F}_n^*(x) - \hat{F}_n(x))$. It readily follows that,

$$P^*[T_n(\hat{F}_n^*) \le t] \stackrel{w}{\Rightarrow} P[\int \psi(x,y)dW^0(F(x))dW^0(F(y)) \tag{2.8}$$
$$+2\int \psi(x,y)dW^0(F(x))d\tilde{W}^0(F(y))$$
$$+\int \psi(x,y)d\tilde{W}^0(F(x))d\tilde{W}^0(F(y)) \le t|\tilde{W}^0]$$

where $\tilde{W}^0$, W^0 are independent Brownian Bridges.

This is again an instance where $\theta(F)$ is well defined for $F \in \mathcal{F}$ but $\theta_n(F)$ doesn't converge for $F \notin \mathcal{F}_0$.

Example 6 Nondifferentiable functions of the empirical: (Beran and Srivastava (1985) and Dümbgen (1993))

Let $\mathcal{F}_0 = \{F : E_F X_1^2 < \infty\}$ and

$$T_n(\hat{F}_n, F) = \sqrt{n}(h(\bar{X}) - h(\mu(F)))$$

when $\mu(F) = E_F X_1$. If h is differentiable the bootstrap distribution of T_n is, of course, consistent. But take $h(x) = |x|$, differentiable everywhere except at 0. It is easy to see then that if $\mu(F) \neq 0$, $\mathcal{L}_n(F) \to \mathcal{N}(0, \mathrm{Var}_F(X_1))$ but if $\mu(F) = 0$, $\mathcal{L}_n(F) \to |\mathcal{N}(0, \mathrm{Var}_F(X_1))|$.

The bootstrap is consistent if $\mu \neq 0$ but not if $\mu = 0$. We can argue as follows. Under $\mu = 0$, $(\sqrt{n}(\bar{X}^* - \bar{X}), \sqrt{n}\bar{X})$ are asymptotically independent $\mathcal{N}(0, \sigma^2(F))$. Call these variable Z and Z'. Then, $\sqrt{n}(|\bar{X}^*| - |\bar{X}|) \stackrel{w}{\Rightarrow} |Z + Z'| - |Z'|$, a variable whose distribution is not the same as that of $|Z|$. The bootstrap distribution, as usual, converges (weakly) to the (random) conditional distribution of $|Z + Z'| - |Z'|$ given Z'. This phenomenon was first observed in a more realistic context by Beran and Srivastava. Dümbgen (1993) constructs similar reasonable though more complicated examples where the bootstrap distribution never converges. If we represent $T_n(\hat{F}_n, F) = \sqrt{n}(T(\hat{F}_n) - T(F))$ in these cases then there is no linear $\dot{T}(F)$ such that $\sqrt{n}(T(\hat{F}_n) - T(F)) \approx \sqrt{n}\dot{T}(F)(\hat{F}_n - F)$ which permits the argument of Bickel-Freedman (1981). □

2.3. **Possible remedies**

Putter and van Zwet (1993) show that if $\theta_n(F)$ is continuous for every n on $\mathcal{F}$ and there is a consistent estimate $\tilde{F}_n$ of F then bootstrapping from $\tilde{F}_n$ will work, i.e. $\theta_n(\tilde{F}_n)$ will be consistent except possibly for F in a "thin" set.

If we review our examples of bootstrap failure, we can see that constructing suitable $\tilde{F}_n \in \mathcal{F}_0$ and consistent is often a remedy that works for all $F \in \mathcal{F}_0$ not simply the complement of a set of the second category. Thus in example 3 taking $\tilde{F}_n$ to be $\hat{F}_n$ kernel smoothed with bandwidth $h_n \to 0$ if $nh_n^2 \to 0$ works. In the first and simplest case of example 4 it is easy to see, Freedman (1981), that taking $\tilde{F}_n$ as the empirical distribution of $X_i - \bar{X}$, $1 \le i \le n$ which has mean 0 and thus belongs to $\mathcal{F}_0$ will work. The appropriate choice of $\tilde{F}_n$ in the other examples of bootstrap failure is less clear. For instance, example 4 calls for $\tilde{F}_n$ with estimated tails of the right order but how to achieve this is not immediate.

A general approach which we believe is worth investigating is to approximate $\mathcal{F}_0$ by a nested sequence of parametric models, (a sieve), $\{\mathcal{F}_{0,m}\}$, and use the M.L.E. $\tilde{\mathcal{F}}_{m(n)}$ for $\mathcal{F}_{0,m(n)}$, for a suitable sequence $m(n) \to \infty$. See Shen andWong (1994) for example.

The alternative approach we study is to change θ_n itself as well as possibly its argument. The changes we consider are the m out of n with replacement bootstrap, the $(n-m)$ out of n jackknife or $\binom{n}{m}$ bootstrap discussed by Wu (1990) and Politis and Romano (1992), and what we call the sample splitting.

3. The m out of n bootstraps

Let h be a bounded real valued function defined on the range of T_n, for instance, $t \to 1(t \le t_0)$.

We view as our goal estimation of $\theta_n(F) \equiv E_F(h(T_n(\hat{F}_n, F)))$. More complicated functionals such as quantiles are governed by the same heuristics and results as those we detail below. Here are the procedures we discuss.

i) *The n/n bootstrap (The nonparametric bootstrap)*

Let,

$$B_n(F) \quad = \quad E^* h(T_n(\hat{F}_n^*, F))$$

$$= \quad n^{-n} \sum_{(i_1,\ldots,i_n)} h(T_n(X_{i_1},\ldots,X_{i_n},F))$$

Then, $B_n \equiv B_n(\hat{F}_n) = \theta_n(\hat{F})$ is the m/n bootstrap.

ii) *The m/n bootstrap*

Let

$$B_{m,n}(F) \equiv n^{-m} \sum_{(i_1,\ldots,i_m)} h(T_m(X_{i_1},\ldots,X_{i_m},F))$$

Then, $B_{m,n} \equiv B_{m/n}(\hat{F}_n) = \theta_m(\hat{F}_n)$ is the m/n bootstrap.

iii) *The $\binom{n}{m}$ bootstrap*

Let

$$J_{m,n}(F) = \binom{n}{m}^{-1} \sum_{i_1<\ldots<i_m} h(T_m(X_{i_1},\ldots,X_{i_m},F))$$

Then, $J_{m,n} \equiv J_{m,n}(\hat{F}_n)$ is the $\binom{n}{m}$ bootstrap.

iv) *Sample splitting*

Suppose $n = mk$. Define,

$$N_{m,n}(F) \equiv k^{-1} \sum_{j=0}^{k-1} h(T_m(X_{jm+1},\ldots,X_{(j+1)m},F))$$

and $N_{m,n} \equiv N_{m,n}(\hat{F}_n)$ is the sample splitting estimates. For safety in practice one should start with a random permutation of the X_i.

The motivation behind $B_{m(n),n}$ for $m(n) \to \infty$ is clear. Since, by (2.1), $\theta_{m(n)}(F) \to \theta(F)$, $\theta_{m(n)}(\hat{F}_n)$ has as good a rationale as $\theta_n(\hat{F}_n)$. To justify J_m note that we can write $\theta_m(F) = \theta_m(\underbrace{F \times \ldots \times F}_{m})$ since it's a parameter of the law of $T_m(X_1,\ldots,X_m,F)$. We now approximate $F \times \ldots \times F$ not by the m dimensional product measure $\underbrace{\hat{F}_n \times \ldots \times \hat{F}_n}_{m}$ but by sampling without replacement. Thus sample splitting is just k fold cross validation and represents a crude approximation to $\underbrace{F \times \ldots \times F}_{m}$.

The sample splitting method requires the least computation of any of the lot. Its obvious disadvantages are that it relies on an arbitrary partition of the sample and that since both m and k should be reasonably large, n has to be really substantial. This method and compromises between it and the $\binom{n}{m}$ bootstrap are studied in Blom (1976) for instance. The $\binom{n}{m}$ bootstrap differs from the m/n by $o_P(1)$ if $m = o(n^{1/2})$. Its advantage is that it never presents us with the ties which make resampling not look like sampling. As a consequence, as we note in theorem 1, it is consistent under really minimal conditions. On the other hand it is somewhat harder to implement by simulation. We shall study both of these methods further below in terms of their accuracy.

A simple and remarkable result on $J_{m(n)}$ has been obtained by Politis and Romano (1992), generalizing Wu (1990). This result was also independently noted and generalized by Götze (1993). Here is a version of the Götze result and its easy proof.

Theorem 1: *Suppose* $\frac{m}{n} \to 0$, $m \to \infty$.

Then,

$$J_m(F) = \theta_m(F) + O_P((\frac{m}{n})^{\frac{1}{2}}). \tag{3.1}$$

If h *is continuous and*

$$T_m(X_1, \ldots, X_m, F) = T_m(X_1, \ldots, X_m, \hat{F}_n) + o_p(1) \tag{3.2}$$

then

$$J_m = \theta_m(F) + o_p(1) \tag{3.3}$$

Proof: Suppose T_m doesn't depend on F. Then, J_m is a U statistic with kernel $h(T_m(x_1, \ldots, x_m))$ and $E_F J_m = \theta_m(F)$ and (3.1) follows immediately. For (3.2) note that

$$E_F|J_m - \binom{n}{m}^{-1} \sum_{i_i < \ldots < i_m} h(T_m(X_{i_1}, \ldots, X_{i_m}, F))| \leq \tag{3.4}$$

$$E_F|h(T_m(X_1, \ldots, X_m, \hat{F}_n)) - h(T_m(X_1, \ldots, X_m, F))|$$

and (3.2) follows by bounded convergence. These results follows in the same way and even more easily for N_m. Note that if T_m doesn't depend on F, $E_F N_m = \theta_m(F)$ and,

$$\mathrm{Var}_F(N_m) = \frac{m}{n}\mathrm{Var}_F(h(T_m(X_1, \ldots, X_m))) > \mathrm{Var}_F(J_m) \quad \Box \tag{3.5}$$

Note: It may be shown, more generally under (3.2), that for example distances between the $\binom{n}{m}$ bootstrap distributionsof $T_m(\hat{F}_m, F)$ and $\mathcal{L}_m(F)$ are also $O_P(\frac{m}{n})^{1/2}$.

Let $X_j^{(i)} = (X_j, \ldots, X_j)_{1\times i}$

$$h_{i_1,\ldots,i_r}(X_1, \ldots, X_r) = \frac{1}{r!} \sum_{1\le j_1 \ne \ldots \ne j_r \le r} h(T_m(X_{j_1}^{(i_1)}, \ldots, X_{j_r}^{(i_r)}, F)). \tag{3.6}$$

For vectors $\mathbf{i} = (i_1, \ldots, i_r)$ in the index set

$$\Lambda_{r,m} = \{(i_1, \ldots, i_r) : 1 \le i_1 \le \ldots \le i_r \le m, i_1 + \ldots + i_r = m\}.$$

Then

$$B_{m,n}(F) = \sum_{r=1}^{m} \sum_{\mathbf{i}\in\Lambda_{r,m}} \omega_{m,n}(\mathbf{i}) \frac{1}{\binom{n}{r}} \sum_{1\le j_1<\ldots<j_r\le m} h_{\mathbf{i}}(X_{j_1}, \ldots, X_{j_r}, F) \tag{3.7}$$

where

$$\omega_{m,n}(\mathbf{i}) = \binom{n}{r}\binom{m}{i_1, \ldots, i_r}/n^m.$$

Let

$$\theta_{m,n}(F) = E_F B_{m,n}(F) = \sum_{r=1}^{m} \sum_{\mathbf{i}\in\Lambda_{r,m}} \omega_{m,n}(\mathbf{i}) E_F h_{\mathbf{i}}(X_1, \ldots, X_r). \tag{3.8}$$

Finally, let

$$\delta_m(\frac{r}{m}) \equiv \max\{|E_F h_{\mathbf{i}}(X_1, \ldots, X_r) - \theta_m(F)| : \mathbf{i} \in \Lambda_{r,m}\} \tag{3.9}$$

and define $\delta_m(x)$ by extrapolation on $[0,1]$. Note that $\delta_m(1) = 0$.

Theorem 2: *Under the conditions of theorem 1*

$$B_{m,n}(F) = \theta_{m,n}(F) + O_P(\frac{m}{n})^{\frac{1}{2}}. \tag{3.10}$$

If further,

$$\delta_m(1 - xm^{-1/2}) \to 0 \tag{3.11}$$

uniformly for $0 \le x \le M$, all $M < \infty$, and $m = o(n)$, then

$$\theta_{m,n}(F) = \theta_m(F) + o(1). \tag{3.12}$$

Finally if,

$$T_m(X_1^{(i_n)}, \ldots, X_r^{(i_r)}, F) = T_m(X_1^{(i_1)}, \ldots, X_r^{(i_r)}, \hat{F}_n) + o_P(1) \tag{3.13}$$

whenever $\mathbf{i} \in \Lambda_{r,m}$, $m \to \infty$ *and* $\max\{i_1, \ldots, i_r\} = O(m^{1/2})$ *then, if* $m \to \infty$, $m = o(n)$,

$$B_m = \theta_m(F) + o_p(1). \tag{3.14}$$

The proof of theorem 2 will be given in the appendix. There too we will sketch that, in the examples we have discussed and some others, $J_{m(n)}$, $B_{m(n)}$, $N_{m(n)}$ are consistent for $m(n) \to \infty$, $\frac{m}{n} \to 0$.

According to theorem 2, if T_n doesn't depend on F the m/n bootstrap works as well as the $\binom{n}{m}$ bootstrap if the value of T_m is not greatly affected by on the order of $\sqrt{m}$ ties in its argument. Some condition is needed. Consider $T_n(X_1, \ldots, X_n) = 1(X_i = X_j$ for some $i \neq j)$ and suppose F is continuous. The $\binom{n}{m}$ bootstrap gives $T_m = 0$ as it should. If $m \neq o(\sqrt{n})$ so that the $\binom{n}{m}$ and m/n bootstraps don't coincide asymptotically the m/n bootstrap gives $T_m = 1$ with positive probability. Finally, (3.13) is the natural extension of (3.2) and is as easy to verify in all our examples.

A number of other results are available for m out of n bootstraps.

Giné and Zinn (1989) have shown quite generally that when $\sqrt{n}(\hat{F}_n - F)$ is viewed as a member of a suitable Banach space $\mathcal{F}$ and,

a) $T_n(X_1, \ldots, X_n, F) = t(\sqrt{n}(\hat{F}_n - F))$ for t continuous

b) $\mathcal{F}$ is not too big

then B_n and $B_{m(n)}$ are consistent.

Praestgaard and Wellner (1993) extend these results to $J_{m(n)}$ with $m = o(n)$. Finally, under the Giné-Zinn conditions,

$$\|\sqrt{m}(\hat{F}_n - F)\| = (\frac{m}{n})\|\sqrt{n}(\hat{F}_n - F)\| = O_P(\frac{m}{n})^{1/2} \tag{3.15}$$

if $m = o(n)$. Therefore,

$$t(\sqrt{m}(\hat{F}_m - \hat{F}_n)) = t(\sqrt{m}(\hat{F}_m - F)) + o_p(1) \tag{3.16}$$

and consistency of N_m if $m = o(n)$ follows from the original Giné-Zinn result.

We close with a theorem on the parametric version of the m/n bootstrap which gives a stronger property than that of theorem 1.

Let $\mathcal{F}_0 = \{F_\theta : \theta \in \Theta \subset R^p\}$ where Θ is open and the model is regular. That is, θ is identifiable, the F_θ have densities f_θ with respect to a σ finite μ and the map $\theta \to \sqrt{f_\theta}$ is continuously Hellinger differentiable with nonsingular derivative. By a result of LeCam, see Bickel, Klaassen, Ritov, Wellner (1992) for instance, there exists an estimate $\hat{\theta}_n$ such that, for all θ,

$$\int (f_{\hat{\theta}_n}^{\frac{1}{2}}(x) - f_\theta^{1/2}(x))^2 d\mu(x) = O_{P_\theta}(\frac{1}{n}). \tag{3.17}$$

Theorem 3: *Suppose $\mathcal{F}_0$ is as above. Let $F_\theta^m \equiv \underbrace{F_\theta \times \ldots \times F_\theta}_{m}$ and $\|\cdot\|$ denote the variational norm. Then*

$$\|F_{\hat{\theta}_n}^m - F_\theta^m\| = O_P((\frac{m}{n})^{1/2}) \tag{3.18}$$

Proof: This is a consequence of the relations (LeCam (1986),

$$\|F_{\theta_0}^m - F_{\theta_1}^m\| \leq H(F_{\theta_0}^m, F_{\theta_1}^m)[(2 - H^2(F_{\theta_0}^m, F_{\theta_1}^m)] \tag{3.19}$$

where

$$H^2(F, G) = \frac{1}{2}\int (\sqrt{dF} - \sqrt{dG})^2 \tag{3.20}$$

and

$$\begin{aligned} H^2(F_{\theta_0}^m, F_{\theta_1}^m) &= 1 - (\textstyle\int \sqrt{f_{\theta_0} f_{\theta_1}} d\mu)^m \\ &= 1 - (1 - H^2(F_{\theta_0}, F))^m. \end{aligned} \tag{3.21}$$

Substituting (3.21) into (3.20) and using (3.17) we obtain

$$\|F_{\hat{\theta}_n}^m - F_\theta^m\| = \tag{3.22}$$

$$O_{P_\theta}(1 - \exp O_{P_\theta}(\frac{m}{n}))^{\frac{1}{2}}(1 + \exp O_{P_\theta}(\frac{m}{n})^{\frac{1}{2}}) = O_{P_\theta}(\frac{m}{n})^{\frac{1}{2}} \qquad \square$$

This result is weaker than theorem 1 since it refers only to the parametric bootstrap. It is stronger since even for $m = 1$ when sampling with and without replacement coincide $\|\hat{F}_n - F_\theta\| = 1$ for all n if F_θ is continous.

4. Performance of B_m, J_m, and N_m as estimates of $\theta_n(F)$

As we have noted if we take $m(n) = o(n)$ then in all examples considered in which B_n is inconsistent, $J_{m(n)}$ $B_{m(n)}$, $N_{m(n)}$ are consistent. Two obvious questions are,

1) How do we choose $m(n)$?

2) Is there a price to be paid for using $J_{m(n)}$, $B_{m(n)}$, or $N_{m(n)}$ when B_n is consistent?

We shall turn to the first very difficult question in a forthcoming paper on diagnostics. The answer to the second is, in general, yes. To make this precise we take the point of view of Beran (1983) and assume that at least on $\mathcal{F}_0$,

$$\theta_n(F) = \theta(F) + \theta'(F)n^{-1/2} + O(n^{-1}) \tag{4.1}$$

where $\theta(F)$ and $\theta'(F)$ are regularly estimable on $\mathcal{F}_0$ in the sense of Bickel, Klaassen, Ritov and Wellner (1993) and $O(n^{-1})$ is uniform on Hellinger compacts. There are a number of general theorems which lead to such expansions. See, for example, Bentkus, Götze and van Zwet (1994).

Somewhat more generally than Beran, we exhibit conditions under which $B_n = \theta_n(\hat{F}_n)$ is fully efficient as an estimate of $\theta_n(F)$ and show that the m out of n bootstrap with $\frac{m}{n} \to 0$ has typically relative efficiency 0.

We formally state a theorem which applies to fairly general parameters θ_n. Suppose ρ is a metric on $\mathcal{F}_0$ such that

$$\rho(\hat{F}_n, F_0) = O_{P_{F_0}}(n^{-1/2}) \text{ for all } F_0 \in \mathcal{F}_0. \tag{4.2}$$

Further suppose

A. $\theta(F)$, $\theta'(F)$ are ρ Fréchet differentiable in $\mathcal{F}$ at $F_0 \in \mathcal{F}_0$. That is,

$$\theta(F) = \theta(F_0) + \int \psi(x, F_0)dF(x) + o(\rho(F, F_0)) \tag{4.3}$$

for $\psi \in L_2^0(F_0) \equiv \{h : \int h^2(x)dF_0(x) < \infty, \int h(x)dF_0(x) = 0\}$ and θ' obeys a similar identity with ψ replaced by another function $\psi' \in L_2^0(F_0)$. Suppose further

B. The tangent space of $\mathcal{F}_0$ at F_0 as defined in Bickel et al. (1993) is $L_2^0(F_0)$ so that ψ and ψ' are the efficient influence functions of θ, θ'. Essentially, we require that in estimating F there is no advantage in knowing $F \in \mathcal{F}_0$.

Finally, we assume,

C. For all $M < \infty$,

$$\sup\{|\theta_m(F) - \theta(F) - \theta'(F)m^{-1/2}| : \rho(F, F_0) \le Mn^{-1/2}, F \in \mathcal{F}\} = O(m^{-1}). \tag{4.4}$$

a strengthened form of (4.1). Then,

Theorem 4: *Under regularity of θ, θ' and A and C at F_0,*

$$\theta_m(\hat{F}_n) \equiv \theta(F_0) + \theta'(F_0)m^{-1/2} + \frac{1}{n}\sum_{i=1}^{n}(\psi(X_i, F_0) + \psi'(X_i, F_0)m^{-1/2}) \tag{4.5}$$

$$+O(m^{-1}) + o_p(n^{-1/2}).$$

If B also holds, $\theta_n(\hat{F}_n)$ is efficient. If in addition, $\theta'(F_0) \neq 0$, and $\frac{m}{n} \to 0$ the efficiency of $\theta_m(\hat{F}_n)$ is 0.

Proof: The expansions of $\theta(\hat{F}_n)$ $\theta'(\hat{F}_n)$ are immediate by Fréchet differentiability and (4.5) follows by plugging these into (4.1). Since θ, θ' are assumed regular, ψ and ψ' are their efficient influence functions. Full efficiency of $\theta_n(\hat{F}_n)$ follows by general theory as given in Beran (1983) for special cases or as extending Theorem 2, p.63 of Bickel et al (1993) in an obvious way. On the other hand, if $\theta'(F_0) \neq 0$, $\sqrt{n}(\theta_m(\hat{F}_n) - \theta_n(F_0))$ has asymptotic bias $(\sqrt{\frac{n}{m}} - 1)\theta'(F_0) + O(\frac{\sqrt{n}}{m}) = \sqrt{\frac{n}{m}}(1 + o(1))\theta'(F_0) \to \pm\infty$ and inefficiency follows. □

Inefficiency results of the same type or worse may be proved about J_m and N_m but require going back to $T_m(X_1, \ldots, X_m, F)$ since J_m and B_n are not related in a simple way. We pursue this only by way of example 1. If $\theta_n(F) = var_F(\sqrt{n}(\bar{X} - \mu(F)) = \theta(F)$, $B_m = B_n$ but,

$$J_m = \sigma^2(\hat{F}_n)(1 - \frac{m-1}{n-1}). \tag{4.6}$$

Thus, since $\theta'(F) = 0$ here, B_m is efficient but J_m has efficiency 0 if $\frac{m}{\sqrt{n}} \to \infty$. N_m evidently behaves in the same way.

It is true that the bootstrap is often used not for estimation but for setting confidence bounds. This is clearly the case for example 1b), the bootstrap t where $\theta(F)$ is known in advance. For example, Efron's percentile bootstrap uses the $(1-\alpha)$th quantile of the bootstrap distribution of $\bar{X}$ as a level $(1-\alpha)$ approximate upper confidence bound for μ. As is well known by now – see Hall (1993), for example, this estimate although, when suitably normalized, efficiently estimating the $(1-\alpha)$th quantile of the distribution of $\sqrt{n}(\bar{X}-\mu)$ does not improve to order $n^{-1/2}$ over the coverage probability of the usual Gaussian based $\bar{X}+z_{1-\alpha}\frac{s}{\sqrt{n}}$. However, the confidence bounds based on the bootstrap distribution of the t statistic $\sqrt{n}(\bar{X}-\mu(F))/s$ get the coverage probability correct to order $n^{-1/2}$. Unfortunately, this advantage is lost if one were to use the $1-\alpha$ quantile of the bootstrap distribution of $T_m(\hat{F}_m, F)=\sqrt{m}(\bar{X}_m-\mu(F))/s_m$ where $\bar{X}_m$ and s_m^2 are the mean and usual estimate of variance based on a sample of size m. The reason is that, in this case, the bootstrap distribution function is;

$$\Phi(t)-m^{-1/2}c(\hat{F}_n)\varphi(t)H_2(t)+O_P(m^{-1}) \tag{4.7}$$

rather than the needed,

$$\Phi(t)-n^{-1/2}c(\hat{F}_n)\varphi(t)H_2(t)+O_P(n^{-1}).$$

The error committed is of order $m^{-1/2}$. More general formal results can be stated but we do not pursue this.

The situation for $J_{m(n)}$ and $N_{m(n)}$ which function under minimal conditions, is even worse as we discuss in the next section.

5. Remedying the deficiencies of $B_{m(n)}$ when B_n is correct: Extrapolation

In Bickel and Yahav (1988), motivated by considerations of computational economy, we considered situations in which θ_n has an expansion of the form (4.1) and proposed using B_m at $m=n_0$ and $m=n_1$, $n_0<n_1<<n$ to produce estimates of θ_n which behave like B_n. We sketch the argument for a special case.

Suppose that, as can be shown for a wide range of situations, if $m\to\infty$,

$$B_m=\theta_m(\hat{F}_n)=\theta(\hat{F}_n)+\theta'(\hat{F}_n)m^{-1/2}+O_P(m^{-1}). \tag{5.1}$$

Then, if $n_1>n_0\to\infty$

$$\theta'(\hat{F}_n)=(B_{n_0}-B_{n_1})(n_0^{-1/2}-n_1^{-1/2})^{-1}+O_P(n_0^{-1/2}) \tag{5.2}$$

$$\theta(\hat{F}_n) = \frac{n_0^{-1/2} B_{n_1} - n_1^{-1/2} B_{n_0}}{n_0^{-1/2} - n_1^{-1/2}} + O_P(n_0^{-1}) \tag{5.3}$$

and hence a reasonable estimate of B_n is,

$$B_{n_0,n_1} \equiv \frac{n_0^{-1/2} B_{n_1} - n_1^{-1/2} B_{n_0}}{n_0^{-1/2} - n_1^{-1/2}} + \frac{(B_{n_0} - B_{n_1})}{n_0^{-1/2} - n_1^{-1/2}} n^{-1/2}.$$

More formally,

Proposition: *Suppose* $\{\theta_m\}$ *obey C of section 4 and* $n_0 n^{-1/2} \to \infty$. *Then,*

$$B_{n_0,n_1} = B_n + o_p(n^{-1/2}). \tag{5.4}$$

Hence, under the conditions of theorem 3 B_{n_0,n_1} *is efficient for estimating* $\theta_n(F)$.

Proof: Under C, (5.4) holds. By construction,

$$\begin{aligned} B_{n_0,n_1} &= \theta(\hat{F}_n) + \theta'(\hat{F}_n) n^{-1/2} + O_P(n_0^{-1}) + O_P(n_0^{-1/2} n^{-1/2}) \\ &= \theta_n(\hat{F}_n) + O_P(n_0^{-1}) + O_P(n_0^{-1/2} n^{-1/2}) + O_P(n^{-1}) = \theta_n(\hat{F}_n) + O_P(n_0^{-1}) \end{aligned} \tag{5.5}$$

and (5.4) follows. □

Assorted variations can be played on this theme depending on what we know or assume about θ_n. If, as in the case where T_n is a t statistic, the leading term $\theta(F)$ in (4.1) is $\equiv \theta_0$ independent of F, estimation of $\theta(F)$ is unnecessary and we need only one value of $m = n_0$. We are led to a simple form of estimate, since ψ of theorem 4 is 0,

$$\hat{\theta}_{n_0} = (1 - (\frac{n_0}{n})^{1/2})\theta_0 + (\frac{n_0}{n})^{\frac{1}{2}} B_{n_0}. \tag{5.6}$$

This kind of interpolation is used to improve theoretically the behaviour of B_{m_0} as an estimate of a parameter of a stable distribution by Hall and Jing (1993) though we argue below that the improvement is somewhat illusory.

If we apply (5.4) to construct a bootstrap confidence bound we expect the coverage probability to be correct to order $n^{-1/2}$ but the error is $O_P((n_0 n)^{-1/2})$ rather than $O_P(n^{-1})$ as with B_n. We do not pursue a formal statement.

5.1. Extrapolation of J_m and N_m

We discuss extrapolation for J_m and N_m only in the context of the simplest example 1, where the essential difficulties become apparent and omit general theorems.

In work in progress, Götze and coworkers are developing expansions for general symmetric statistics under sampling from a finite population. These results will permit general statements of the same qualitative nature as our discussion of example 1. Consider $\theta_m(F) = P_F[\sqrt{m}(\bar{X}_m - \mu(F)) \le t]$. If $EX_1^4 < \infty$ and the X_i obey Cramér's condition,

$$\theta_m(F) = \Phi(\frac{t}{\sigma(F)}) - K_3(F)\frac{\varphi}{6\sqrt{m}}(\frac{t}{\sigma(F)})H_2(\frac{t}{\sigma(F)}) + O(m^{-1}), \tag{5.7}$$

where $\sigma^2(F)$ and $K_3(F)$ are the second and third cumulants of F. By Singh (1981), $B_m = \theta_m(\hat{F}_n)$ has the same expansion with F replaced by $\hat{F}_n$. However, by an easy extension of results of Robinson (1978) and Babu and Singh (1985),

$$J_m = \Phi(\frac{t}{\hat{K}_2 m}) - \varphi(\frac{t}{\hat{K}_{2m}^{1/2}})\frac{\hat{K}_{3m}}{6m^{1/2}}H_2(\frac{t}{\hat{K}_{2m}^{1/2}}) + O_P(m^{-1}) \tag{5.8}$$

where

$$\hat{K}_{2m} = \sigma^2(\hat{F}_n)(1 - \frac{m-1}{n-1}) \tag{5.9}$$

$$\hat{K}_{3m} = K_3(\hat{F}_n)(1 - \frac{m-1}{n-1})(1 - \frac{2(m-1)}{n-2}). \tag{5.10}$$

The essential character of expansion (5.8), if $\frac{m}{n} = o(1)$, is

$$J_m = \theta(\hat{F}_n) + m^{-1/2}\theta'(\hat{F}_n) + \frac{m}{n}\gamma_n + O_P(m^{-1} + (\frac{m}{n})^2 + \frac{m^{\frac{1}{2}}}{n}) \tag{5.11}$$

where γ_n is $O_P(1)$ and independent of m. The $\frac{m}{n}$ terms essentially come from the finite population correction to the variance and higher order cumulants of means of samples from a finite population. They reflect the obvious fact that if $\frac{m}{n} \to \lambda > 0$, J_m is, in general, incorrect even to first order. For instance, the variance of the $\binom{n}{m}$ bootstrap distribution corresponding to $\sqrt{m}(\bar{X} - \mu(F))$ is $\frac{1}{n}\sum(X_i - \bar{X})^2(1 - \frac{m-1}{n-1})$ which converges to $\sigma^2(F)(1-\lambda)$ if $\frac{m}{n} \to \lambda > 0$. What this means is that if expansions (4.1), (5.1) and (5.11) are valid, then using $J_{m(n)}$ again gives efficiency 0 compared to B_n. Worse is that (5.2) with J_{n_0}, J_{n_1} replacing B_{n_0}, B_{n_1} will not work since the $\frac{n_1}{n}$ terms remain and make a contribution larger than $n^{-1/2}$ if $\frac{n_1}{n^{1/2}} \to \infty$. Essentially it is necessary to estimate the coefficient of $\frac{m}{n}$ and remove the contribution of this term also while keeping the three required values of m: $n_0 < n_1 < n_2$ such that the error $O(\frac{1}{n_0} + (\frac{n_2}{n})^2)$ is $o(n^{-1/2})$. This essentially means that n_0, n_1, n_2 have order larger than $n^{1/2}$ and smaller that $n^{3/4}$.

This effect persists if we seek to use an extrapolation of J_m for the t statistic. The coefficient of $\frac{m}{n}$ as well as $m^{-1/2}$ needs to be estimated. An alternative here and perhaps more generally is to modify the t statistic being bootstrapped and extrapolated. Thus $T_m(X_1, \ldots, X_m, F) \equiv \sqrt{m}\frac{(\bar{X}_m - \mu(F))}{\hat{\sigma}(1-\frac{m-1}{n-1})^{\frac{1}{2}}}$ leads to an expansion for J_m of the form,

$$J_m = \Phi(t) + \theta'(\hat{F}_n)m^{-1/2} + O_P(m^{-1} + m/n), \tag{5.12}$$

and we again get correct coverage to order $n^{-1/2}$ by fitting the $m^{-1/2}$ term's coefficient, weighting it by $n^{-1/2} - m^{-1/2}$ and adding it to J_m.

If we know, as we sometimes at least suspect in symmetric cases, that $\theta(F) = 0$, we should appropriately extrapolate linearly in m^{-1} rather than $m^{-1/2}$.

The sample splitting situation is less satisfactory in the same example. Under (5.1), the coefficient of $\frac{1}{\sqrt{m}}$ is asymptotically constant. Put another way, the asymptotic correlation of B_m, $B_{\lambda m}$ as $m, n \to \infty$ for fixed $\lambda > 0$ is 1. This is also true for J_m under (5.11). However, consider N_m and N_{2m} (say) if $T_m = \sqrt{m}(\bar{X}_m - \mu(F))$. Let h be continuously boundedly differentiable, $n = 2km$. Then

$$\text{cov}(N_m, N_{2m}) = \frac{1}{k}\text{cov}(h(m^{-1/2}(\sum_{j=1}^{m}(X_j - \bar{X}))), h((2m)^{-1/2}\sum_{j=1}^{2m}(X_j - \bar{X}))). \tag{5.13}$$

Thus, by the central limit theorem,

$$\text{corr}(N_m, N_{2m}) \to \frac{1}{2}\frac{\text{cov}}{\text{var}(Z_1)}(h(Z_1), h\frac{(Z_1 + Z_2)}{\sqrt{2}}) \tag{5.14}$$

where Z_1, Z_2 are independent Gaussian $\mathcal{N}(0, \sigma^2(F))$ and $\sigma^2(F) = \text{Var}_F(X_1)$. More generally, viewed as a process in m for fixed n, N_m centered and normalized is converging weakly to a non degenerate process. Thus, extrapolation doesn't make sense for N_m.

Two questions naturally present themselves.

(a) How do these games play out in practice rather than theory?

(b) If the expansions (5.1) and (5.11) are invalid beyond the 0th order, the usual situation when the nonparametric bootstrap is inconsistent, what price do we pay theoretically for extrapolation?

Simulations giving limited encouragement in response to question (a) are given in Bickel and Yahav (1988). We give some further evidence in section 7. We turn to question (b) in the next section.

6. Behaviour of the smaller resample schemes when B_n is inconsistent, and alternatives

The class of situations in which B_n does not work is too poorly defined for us to come to definitive conclusions. But consideration of the examples suggests the following,

A. When, as in example 6, $\theta(F)$, $\theta'(F)$ are well defined and regularly estimable on $\mathcal{F}_0$ we should still be able to use extrapolation (suitably applied) to B_m and possibly to J_m to produce better estimates of $\theta_n(F)$.

B. When, as in all our other examples of inconsistency, $\theta(F)$ is not regularly estimable on $\mathcal{F}_0$ extrapolation shouldn't improve over the behaviour of B_{n_0}, B_{n_1}

C. If n_0, n_1 are comparable extrapolation shouldn't do particularly worse either.

D. A closer analysis of T_n and the goals of the bootstrap may, in these "irregular" cases, be used to obtain procedures which should do better than the m/n or $\binom{n}{m}$ or extrapolation bootstraps.

The only one of these claims which can be made general is C.

Proposition 1: *Suppose*

$$B_{n_1} - \theta_n(F) \asymp B_{n_0} - \theta_n(F) \tag{6.1}$$

where $\asymp$ indicates that the ratio tends to 1. Then, if $\frac{n_0}{n_1} \not\to 1$

$$B_{n_0,n_1} - \theta_n(F) \asymp B_{n_0} - \theta_n(F). \tag{6.2}$$

Proof: Evidently, $\frac{B_{n_0}+B_{n_1}}{2} = \theta_n(F) + \Omega(\epsilon_n)$ where $\Omega(\epsilon_n)$ means that the exact order of the remainder is ϵ_n. On the other hand,

$$\frac{B_{n_0} - B_{n_1}}{n_0^{-1/2} - n_1^{-1/2}}\left(\frac{1}{\sqrt{n}} - \frac{1}{2}\left(\frac{1}{\sqrt{n_0}} + \frac{1}{\sqrt{n_1}}\right)\right)$$

$$= \Omega(\epsilon_n)\left(\sqrt{\frac{n_0}{n}} + \Omega(1)\right)$$

and the proposition follows. □

We illustrate the other three claims in going through the examples.

Example 3: Here, $F^{-1}(0) = 0$,

$$\theta_n(F) = e^{f(0)t}(1 + n^{-1}f'(0))\frac{t^2}{2}) + O(n^{-2}) \tag{6.3}$$

which is of the form (5.1). But the functional $\theta(F)$ is not regular and only estimable at rate $n^{-1/3}$ if one puts a first order Lipschitz condition on $F \in \mathcal{F}_0$. On the other hand,

$$\begin{aligned} \log B_m &= m\log(1 - \hat{F}_n(\frac{t}{m})) = m\log(1 - (\hat{F}_n(\frac{t}{m}) - \hat{F}_n(0))) \qquad (6.4)\\ &= -m(F(\frac{t}{m}) - F(0)) - \frac{m}{\sqrt{n}}\sqrt{n}(\hat{F}_n(\frac{t}{m}) - F(\frac{t}{m}))\\ &\quad + O_P(m(\hat{F}_n(\frac{t}{m}) - F(\frac{t}{m}))^2)\\ &= tf(0) + \Omega(\frac{1}{m}) + \Omega_P(\sqrt{\frac{m}{n}}) + O_P(\frac{1}{n}) \end{aligned}$$

where as before Ω, Ω_P indicate exact order. As Politis and Romano point out, $m = \Omega(n^{1/3})$ yields the optimal rate $n^{-1/3}$ (under f Lipschitz). Extrapolation doesn't help because the $\sqrt{\frac{m}{n}}$ term is not of the form $\gamma_n\sqrt{\frac{m}{n}}$ where γ_n is independent of m. On the contrary, as a process in m, $\sqrt{mn}(\hat{F}_n(\frac{t}{m}) - F(\frac{t}{m}))$ behaves likes the sample path of a stationary Gaussian process. So conclusion B holds in this case.

Example 4: A major difficulty here is defining $\mathcal{F}_0$ narrowly enough so that it is meaningful to talk about expansions of $\theta_n(F)$, $B_n(F)$ etc. If $\mathcal{F}_0$ in these examples is all distributions in the domain of attraction of stable laws or extreme value distributions it is easy to see that $\theta_n(F)$ can converge to $\theta(F)$ arbitrarily slowly. This is even true in example 1 if we remove the Lipschitz condition on f. By putting on conditions as in example 1, it is possible to obtain rates. Hall and Jing (1993) specify a possible family for the stable law attraction domain estimation of the mean mentioned in example 4 in which $B_n = \Omega(n^{-\frac{1}{\alpha}})$ where α is the index of the stable law and α and the scales of the (assumed symmetric) stable distribution are not regularly estimable but for which rates such as $n^{-2/5}$ or a little better are possible. The expansions for $\theta_n(F)$ are not in powers of $n^{-1/2}$ and the expansion for B_n is even more complex. It seems evident that extrapolation doesn't help. Hall and Jing theoretical results and simulations show that $B_{m(n)}$ though consistent, if $m(n)/n \to 0$, is a very poor estimate of $\theta_n(F)$. They obtain at least theoretically superior results by using interpolation between B_m and the, "known up to the value of the stable law index α", value of $\theta(F)$. However, the conditions defining $\mathcal{F}_0$ which permit them to deduce the order of B_n are uncheckable so that this improvement appears illusory.

Example 6: The discontinuity of $\theta(F)$ at $\mu(F) = 0$ under any reasonable specification of $\mathcal{F}_0$ makes it clear that extrapolation cannot succeed. The discontinuity in $\theta(F)$ persists even if we assume $\mathcal{F}_0 = \{\mathcal{N}(\mu, 1) : \mu \in R\}$ and use the parametric bootstrap. In the parametric case it is possible to obtain constant level confidence bounds by inverting the tests for $H : |\mu| = |\mu_0|$ vs $K : |\mu| > |\mu_0|$ using the noncentral χ_1^2 distribution of $(\sqrt{n}\bar{X})^2$. Asymptotically conservative confidence bounds can be constructed in the nonparametric case by forming a bootstrap confidence interval for $\mu(F)$ using $\bar{X}$ and then taking the image of this interval into $\mu \to |\mu|$. So this example illustrates points B and D.

We shall discuss claims A and D in the context of example 5 or rather its simplest case with $T_n(\hat{F}_n, F) = n\bar{X}^2$. We begin with,

Proposition 2: *Suppose $E_F X_1^4 < \infty$, $E_F X_1 = 0$, and F satisfies Cramér's condition. Then,*

$$B_m \equiv P^*[|\sqrt{m}\bar{X}]^2 \leq t^2] = 2\Phi(\frac{t}{\hat{\sigma}}) - 1 + \varphi'(\frac{t}{\hat{\sigma}})m\frac{\bar{X}^2}{\hat{\sigma}^2} \tag{6.5}$$
$$-\frac{\hat{K}_3}{6}\varphi H_2(\frac{t}{\hat{\sigma}})\bar{X} + O_P(\frac{m}{n^{3/2}}) + O_P(\frac{m^2}{n^2}) + O_P(m^{-1}).$$

If $m = \Omega(n^{1/2})$ then

$$P^*[|\sqrt{m}\bar{X}|^2 \leq t^2] = P_F[n\bar{X}^2 \leq t] + O_P(n^{-1/4}) \tag{6.6}$$

and no better choice of $\{m(n)\}$ is possible. If $n_0 < n_1$, $n_0 n^{-1/2} \to \infty$, $n_1 = o(n^{3/4})$,

$$B^{n_0,n_1} \equiv B_{n_0} - n_0\{(B_{n_1} - B_{n_0})/(n_1 - n_0)\} = P_F[n\bar{X}^2 \leq t] + O_P(n^{-1/2}). \tag{6.7}$$

Proof: We make a standard application of Singh (1981). If $\hat{\sigma}^2 \equiv \frac{1}{n}\sum(X_i - \bar{X})^2$, $\hat{K}_3 \equiv \frac{1}{n}\sum(X_i - \bar{X})^3$ we get, after some algebra.

$$P^*[m\bar{X}_m^2 \leq t^2] = 2\Phi(\frac{t}{\hat{\sigma}}) - 1 + \frac{\varphi'}{2}(\frac{t}{\hat{\sigma}})m\bar{X}^2 - \frac{\hat{K}_3}{6}[\varphi H_2](\frac{t}{\hat{\sigma}})\bar{X} \tag{6.8}$$

$$+O_P(\frac{m^2}{n^2}) + O_P(\frac{m}{n^{3/2}}) + O_P(m^{-1})$$

and (6.5) follows. Since $m\bar{X}^2 = \Omega_P(\frac{m}{n})$, (6.6) follows. Finally, from (6.5), if $n_0 n^{-1/2}$, $n_1 n^{-1/2} \to \infty$

$$B_{n_0} - n_0\{(B_{n_1} - B_{n_0})/(n_1 - n_0)\} = 2\Phi(\frac{t}{\hat{\sigma}}) - 1 \tag{6.9}$$

$$-\frac{K_3}{6}\varphi H_2(\frac{t}{\hat{\sigma}})\bar{X} + O_P(n^{-3/4}) + O_P(n^{-1/2}) + O_P(n^{-1/2}).$$

Since $\bar{X} = O_P(n^{-1/2})$, (6.7) follows. □

Example 5: As we noted the case $T_n(\hat{F}_n, F) = n\bar{X}^2$ is the prototype of the use of the m/n bootstrap for testing discussed in Bickel and Ren (1995). ¿From (6.7) of proposition 2 it is clear that extrapolation helps. However, it is not true that B^{n_0,n_1} is efficient since it has an unnecessary component of variance $\frac{\hat{K}_3}{6}[\varphi H_2](\frac{t}{\hat{\sigma}})\bar{X}$ which is negligible only if $K_3(F) = 0$. On the other hand it is easy to see that efficient estimation can be achieved by resampling not the X_i but the residuals $X_i - \bar{X}$, that is, a consistent estimate of F belonging to $\mathcal{F}_0$. So this example illustrates both A and D. Or in the general U or V statistic case, bootstrapping not $T_m(\hat{F}_n, F) \equiv n\int \psi(x,y)d\hat{F}_n(x)d\hat{F}_n(y)$ but rather $n\int \psi(x,y)d(\hat{F}_n - F)(x)d(\hat{F}_n - F)(y)$ is the right thing to do.

7. Simulations and Conclusions

The simulation algorithms were written and carried out by Adele Cutler and Jiming Jiang. Two situations were simulated, one already studied in Bickel and Yahav (1988) where the bootstrap is consistent (essentially example 1) the other (essentially example 3) where the bootstrap is inconsistent.

Sample sizes: $n = 50, 100, 400$

Bootstrap sample sizes: $B = 500$

Simulation size: $N = 2000$

Distributions: **Example** 1: $F = \chi^2_1$; **Example** 3: $F = \chi^2_2$

Statistics:

Example 1a) modified: $T_m^{(a)} = \sqrt{m}(\sqrt{\bar{X}_m} - \sqrt{\mu(F)})$

Example 1b): $T_m^{(b)} = \sqrt{m}\frac{(\bar{X}_m - \mu(F))}{s_m}$ where $s_m^2 = \frac{1}{m-1}\sum_{i=1}^m (X_i - \bar{X}_m)^2$.

Example 3: $T_m^{(c)} = m(\min(X_1, \ldots, X_m) - F^{-1}(0))$

Parameters of resampling distributions: $G_m^{-1}(.1)$, $G_m^{-1}(.9)$ where G_m is the distribution of T_m under the appropriate resampling scheme. We use B, J, N to distinguish the schemes m/n, $\binom{n}{m}$ and sample splitting respectively.

In example 1 the G_m^{-1} parameters were used to form upper and lower "90%" confidence bounds for $\theta \equiv \sqrt{\mu(F)}$. Thus, from $T_m^{(a)}$,

$$\bar{\theta}_{mB} = \sqrt{\bar{X}_n} - \frac{1}{\sqrt{n}} G_{mB}^{-1}(.1) \tag{6.1}$$

for the "90%" upper confidence bound based on the m/n bootstrap and, from $T_m^{(b)}$,

$$\bar{\theta}_{mB} = ((\bar{X}_n - \frac{s_n}{\sqrt{n}} G_{mB}^{-1}(.1))_+)^{1/2} \tag{6.2}$$

where G_{mB} now corresponds to the t statistic. $\underline{\theta}_{mB}$, is defined similarly. The $\bar{\underline{\theta}}_{mJ}$ bounds are defined with G_{mJ} replacing G_{mB}. The $\bar{\underline{\theta}}_{mN}$ bounds are considered only for the unambiguous case m divides n and α an integer multiple of $\frac{m}{n}$. Thus if $m = \frac{n}{10}$, $G_{mN}^{-1}(.1)$ is simply the smallest of the 10 possible values $\{T_m(X_{jm+1}, \ldots, X_{(j+1)m}, \hat{F}_n),\ 0 \le j \le 9\}$.

We also specify 2 subsample sizes $n_0 < n_1$ for the extrapolation bounds, $\underline{\theta}_{n_0,n_1}$ $\bar{\theta}_{n_0,n_1}$. These are defined for $T_m^{(a)}$, for example, by,

(6.3)

$$\bar{\theta}_{n_0,n_1} = \sqrt{\bar{X}_n} - \frac{1}{\sqrt{n}}\{\frac{(G_{n_0B}^{-1}(.1) + G_{n_1B}^{-1}(.1))}{2} + (n^{-1/2} - \frac{1}{2}(n_0^{-1/2} + n_1^{-1/2}))(G_{n_0B}^{-1}(.1) - G_{n_1B}^{-1}(.1)/(n_0^{-1/2} - n_1^{-1/2}).$$

We consider roughly, $n_0 = 2\sqrt{n}$, $n_1 = 4\sqrt{n}$ and specifically, the triples (n, n_0, n_1) : $(50, 15, 30), (100, 20, 40)$ and $(400, 40, 80)$.

In example 3, we similarly study the lower confidence bound on $\theta = F^{-1}(0)$ given by,

$$\underline{\theta}_m = \max(X_1, \ldots, X_n) - \frac{1}{n} G_{mB}^{-1}(.9). \tag{6.4}$$

and the extrapolation lower confidence bound

$$\underline{\theta}_{n_0,n_1} = \min(X_1, \ldots, X_n) - \frac{1}{n}\frac{(G_{n_0B}^{-1}(.9) + G_{n_1B}^{-1}(.9))}{2} \tag{6.5}$$

$$+(n^{-1} - \frac{(n_0^{-1} + n_1^{-1})}{2})(G_{n_0B}^{-1}(.9) - G_{n_1B}^{-1}(.9))(n_0^{-1} - n_1^{-1}).$$

Note that we are using $\frac{1}{m}$ rather than $\frac{1}{\sqrt{m}}$ for extrapolation.

Measures of performance:

$CP \equiv$ Coverage probability, the actual probability under the situation simulated that the region prescribed by the confidence bound covers the true value of the parameter being estimated.

$$RMSE = \sqrt{E(\text{Bound} - \text{Actual quantile bound})^2}.$$

Here the actual quantile bound refers to what we would use if we knew the distribution of $T_n(X_1, \ldots, X_n, F)$. For example for $T_m^{(a)}$ we would replace $G_{mB}^{-1}(.1)$ in (6.1) for $F = \chi_1^2$ by the .1 quantile of the distribution of $\sqrt{n}(\sqrt{\frac{S_m}{m}} - 1)$ where S_m has a χ_m^2 distribution, call it $G_m^{*-1}(.1)$. Thus, here,

$$MSE = \frac{1}{n} E(G_{mB}^{-1}(.1) - G_m^{*-1}(.1))^2.$$

We give in table 1 results for the B_{n_1}, B_n, and B_{n_0,n_1} bounds, based on $T_m^{(b)}$. The $T_m^{(a)}$ bootstrap, as in Bickel-Yahav (1988), has CP and $RMSE$ for B_n, B_{n_0,n_1} and B_{n_1} agreeing to the accuracy of the Monte Carlo and we omit these tables.

We give the corresponding results for lower confidence bounds based on $T_m^{(c)}$ in table 2. Table 3 presents results for sample splitting for $T_m^{(a)}$. Table 4 presents $T_m^{(a)}$ results for the $\binom{n}{m}$ bootstrap.

Table 1: The t bootstrap: example 1b) at 90% nominal level

n		Coverage probabilities (CP) B	B1	BR	RMSE B	B1	BR
50							
	UB	.88	.90	.88	.19	.21	.19
	LB	.90	.90	.90	.15	.15	.15
100							
	UB	.90	.93	.89	.13	.14	.12
	LB	.91	.90	.91	.11	.10	.11
400							
	UB	.91	.94	.90	.06	.07	.06
	LB	.91	.90	.91	.05	.05	.05

Notes a) B1 corresponds to (6.2) or its LCB analogue for $m = n_1(n) = 30, 40, 80$. Similarly B corresponds to $m = n$.

b) BR corresponds to (6.3) or its LCB analogue with $(n_0, n_1) = (15, 30), (20, 40), (40, 80)$.

Table 2: The min statistic bootstrap: example 3 at the nominal 90% level

n		*CP*	*RMSE*
50			
	B	.75	.01
	B1	.78	.07
	BR	.70	.07
	B1S	.82	.07
	BRS	.80	.07

n		*CP*	*RMSE*
100			
	B	.75	.04
	B1	.82	.03
	BR	.76	.04
	B1S	.87	.03
	BRS	.86	.03
400			
	B	.75	.09
	B1	.86	.01
	BR	.83	.01

Notes: a) B corresponds to (6.4) with $m = n$, B1 with $m = n_1 = 30, 40, 80$, B1S with $m = n_1 = 16$.

b) BR corresponds to (6.5) with $(n_0, n_1) = (15, 30), (20, 40), (40, 80)$, BRS with $(n_0, n_1) = (4, 16)$.

Table 3: Sample splitting in example 1a)

		CP		*RMSE*	
n		N	$B_{m(n)}$	N	$B_{m(n)}$
50					
	UB	.82	.86	.32	.18
	LB	.86	.91	.28	.16
100					
	UB	.86	.89	.30	.14
	LB	.84	.90	.26	.12
400					
	UB	.85	.89	.28	.08
	LB	.86	.91	.27	.09

Note: N here refers to $m = .1n$ and $\alpha = .1$

Table 4: The $\binom{n}{m}$ bootstrap and the m/n bootstrap in example 1a)

		CP		*E(Length)*	
n	m	J	B	J	B
50	16	.82	.88	.07	.09
100	16	.86	.88	.04	.05
400	40	.88	.90	.01	.01

Note: These figures are for simulation sizes of $N = 500$ and for 90% confidence intervals. Thus, the end points of the intervals are given by (6.1) and its UCB counterpart for B and J but with .1 replaced by .05. Similarly, $[E(\text{Bound} - \text{Actual quantile bound})^2]^{1/2}$ is replaced by the expected length of the confidence interval.

Conclusions: The conclusions we draw are limited by the range of our simulations. We opted for realistic sample sizes, of 50, 100 and a less realistic 400. For $n = 50$, 100 the subsample sizes $n_1 = 30$ (for $n = 50$) and 40 (for $n = 100$) are of the order $n/2$ rather than $o(n)$. For all sample sizes $n_0 = 2\sqrt{n}$ is not really "of larger order than $\sqrt{n}$". The simulations in fact show the asymptotics as very good when the bootstrap works even for relatively small sample sizes. The story when the bootstrap doesn't work is less clear.

When the bootstrap works (**Example** 1)

- BR and B are very close both in terms of CP, and RMSE even for $n = 50$ from table 1.

- B1's CP though sometimes better than B's consistently differs more from B's and its RMSE follows suit In particular, for UB in table 1, the RMSE of B1 is generally larger. LB exhibits less differences but this reflects that UB is governed by the behaviour of χ_1^2 at 0. In simulations we do not present we get similar sharper differences for LB when F is a heavy tailed distribution such as Pareto with $EX^5 = \infty$.

- The effects, however, are much smaller than we expected. This reflects that these are corrections to the coefficient of the $n^{-1/2}$ term in the expansion. Perhaps the most surprising aspect of these tables is how well B1 performs.

- From table 3 we see that because the m we are forced to by the level considered is small, CP for the sample splitting bounds differs from the nominal level. If $n \to \infty$, $\frac{m}{n} \to .1$ the coverage probability doesn't tend to .1 since the estimated quantile doesn't tend to the actual quantile and both CP and RMSE behave badly compared to B_m. This naive method can be fixed up – see Blom (1976) for instance. However, its simplicity is lost and the $\binom{n}{m}$ or m/n bootstrap seem preferable.

- The $\binom{n}{m}$ bounds are inferior as table 4 shows. This reflects the presence of the finite population correction $\frac{m}{n}$, even though these bounds were considered for the more favorable sample size $m = 16$ for $n = 50$, 100 rather than $m = 30$, 40. Corrections such as those of Bertail (1994) or simply applying the finite population correction to s would probably bring performance up to that of B_{n_1}. But the added complication doesn't seem worthwhile.

When the bootstrap doesn't work (**Example** 3)

- From table 2, as expected, the CP of the n/n bootstrap for the lower confidence bound was poor for all n. For $n_0 = 2\sqrt{n}$, $n_1 = 4\sqrt{n}$, CP for B1 was constantly better than B for all n. BR is worse than B1 but improves with n and was nearly as good as B1 for $n = 400$. For small n_0, n_1 both B1 and BR do much better. However, it is clear that the smaller m of B1S is better than all other choices.

We did not give results for the upper confidence bound because the granularity of the bootstrap distribution of $\min_i X_i$ for these values of m and n made $CP = 1$ in all cases.

Evidently, n_0, n_1 play a critical role here. What apparently is happening is that for n_0, n_1 not sufficiently small compared with n extrapolation picks up the wrong slope and moves the not so good B1 bound even further towards the poor B bound.

A message of these simulations to us is that extrapolation of the B_m plot may carry risks not fully revealed by the asymptotics. On the other hand, if n_0 and n_1 are chosen in a reasonable fashion extrapolation on the $\sqrt{n}$ scale works well when the bootstrap does. Two notes, based on simulations we do not present, should be added to the optimism of

Bickel, Yahav (1988) however. There may be risk if n_0 is really small compared to $\sqrt{n}$. We obtained poor results for BR for the t statistics for $n_0 = 4$ and 2. Thus $n_0 = 4$, $n_1 = 16$ gave the wrong slope to the extrapolation which tended to overshoot badly. Also, taking n_1 and n_0 close to each other, as the theory of the 1988 paper suggests is appropriate for statistics possessing high order expansions when the expansion coefficients are deterministic, gives poor results. It can also be seen theoretically that the sampling variability of the bootstrap for m of the order $\sqrt{n}$ makes this prescription unreasonable.

The principal message we draw is that it is necessary to develop data driven methods of selection of m which lead to reasonable results over situations where both the bootstrap works and where it doesn't. Such methods are being pursued.

Appendix

Proof of Theorem 2: For $\mathbf{i} = (i_1, \ldots, i_r) \in \Lambda_{r,m}$ let $U(\mathbf{i}) = \frac{1}{\binom{n}{r}} \sum\{h_{\mathbf{i}}(X_{j_1}, \ldots, X_{j_r}, F) : 1 \le j_1 < \ldots < j_r \le n\}$. Then, since $h_{\mathbf{i}}$ as defined is symmetric in its arguments it is a U statistic and $\|h\|_\infty$ is an upper bound to its kernel. Hence

(a) $\mathrm{Var}_F U(\mathbf{i}) \le \|h\|_\infty^2 \frac{r}{n}$

On the other hand,

(b) $EU(\mathbf{i}) = E_F h_{\mathbf{i}}(X_1, \ldots, X_r, F)$

and

(c) $B_{m,n}(F) = \sum_{r=1}^m \sum\{w_{m,n}(\mathbf{i}) U(\mathbf{i}) : \mathbf{i} \in \Lambda_{r,m}\}$

by (3.7).

Thus, by (c),

(d) $\mathrm{Var}_F^{\frac{1}{2}} B_{m,n}(F) \le \sum_{r=1}^m \sum \ w_{m,n}(\mathbf{i}) \mathrm{Var}_F^{\frac{1}{2}} U(\mathbf{i}) : \mathbf{i} \in \Lambda_{r,m}\}$

$$\le \max \mathrm{Var}_F^{\frac{1}{2}} U(\mathbf{i}) \le \|h\|_\infty (\frac{m}{n})^{\frac{1}{2}}$$

by (a). This completes the proof of (3.10).

The proof of (3.11) is more involved. By (3.8)

$$(e) \qquad |\theta_{m,n}(F)-\theta(F)| \le \sum_{r=1}^{m}\sum\{|E_F h_{\mathbf{i}}(X_1,\ldots,X_r)-\theta_m(F)|w_{m,n}(\mathbf{i}) : \mathbf{i}\in\Lambda_{r.m}\}.$$

Let,

$$(f) \qquad P_{m,n}[R_m=r]=\sum\{w_{m,n}(\mathbf{i}):\mathbf{i}\in\Lambda_{r,m}\}$$

Expression (f) is easily recognized as the probability of getting $n-r$ empty cells when throwing n balls independently into m boxes without restrictions — see Feller (1968) p.19. Then it is well known or easily seen that

(g) $E_{m,n}(R_m)=n(1-(1-\frac{1}{n})^m)$

(h) $\mathrm{Var}_{m,n}(R_m)=n\{(1-\frac{1}{n})^m-(1-\frac{2}{n})^m\}+n^2\{(1-\frac{2}{n})^m-(1-\frac{1}{n})^{2m}\}$.

It is easy to check that, if $m=o(n)$

(i) $E_{m,n}(R_m)=m(1+O(\frac{m}{n}))$

(j) $\mathrm{Var}_{m,n}(R_m)=O(m)$

so that,

(k) $\frac{R_m}{m}=1+O_P(m^{-1/2})$.

¿From (e),

$$(l) \qquad |\theta_{m,n}(F)-\theta(F)|\le\sum_{r=1}^{m}\delta_m(\frac{r}{m})P_{m,n}[R_m=r].$$

By (k), (l) and the dominated convergence theorem (3.12) follows from (3.11) and (k).

Finally, as in theorem 1, we bound, as in (3.4),

$$(m) \quad |B_{m,n}(F)-B_m(F)|\le\sum_{r=1}^{m}\sum\{E_F|h_{\mathbf{i}}(X_1,\ldots,X_r)-h_{\mathbf{i}}(X_1,\ldots,X_r,\hat F_n)|:\mathbf{i}\in\Lambda_{r,m}\}w_{m,n}(\mathbf{i})$$

where

$$(n) \qquad h_{\mathbf{i}}(X_1,\ldots,X_r,\hat F_n)=\frac{1}{r!}\sum_{1\le j_1\neq\ldots\neq j_r\le r} h(T_m(X_{j_1}^{(i_1)},\ldots,X_{j_r}^{(i_r)},\hat F_n))$$

Let R_m be distributed according to (f) and given $R_m = r$, $(I_1, \ldots, I_r)$ be uniformly distributed on the set of partitions of m into r ordered integers, $I_1 \leq I_2 \leq \ldots \leq I_r$. Then, from (m) we can write

$$(o) \qquad |B_{m,n}(F) - B_m(F)| \leq E\Delta(I_1, \ldots, I_{R_m})$$

where $\|\Delta\|_\infty \leq \|h\|_\infty$. Further, by the continuity of h and (3.13), since $I_1 \leq \ldots \leq I_{R_m}$,

$$(p) \qquad \Delta(I_1, \ldots, I_{R_m}) 1(I_{R_m} \leq \epsilon_m m) \xrightarrow{P} 0$$

whenever $\epsilon_m = O(m^{-1/2})$. Now, $I_{R_m} > \epsilon_m m$,

$$(q) \qquad m = \sum_{j=1}^{R_m} I_j$$

and $I_j \geq 1$ imply that,

$$(r) \qquad m(1 - \epsilon_m) \geq \sum_{j=1}^{R_m - 1} I_j \geq (R_m - 1)$$

Thus,

$$(s) \qquad P_{m,n}(I_{R_m} > \epsilon_m m) \leq P_{m,n}(\frac{R_m}{m} - 1 \leq -\epsilon_m + O(m^{-1})) \rightarrow 0$$

if $\epsilon_m m^{1/2} \rightarrow \infty$. Combining (s), (k) and (p) we conclude that

$$(t) \qquad E\Delta(I_1, \ldots, I_{R_m}) \rightarrow 0$$

and hence (o) implies (3.14). □

The corollary follows from (e) and (f). □

Note that this implies that the m/n bootstrap works if about $\sqrt{m}$ ties do not affect the value of T_m much.

Checking that J_m, B_m, N_m $m = o(n)$ works

The arguments we give for B_m also work for J_m only more easily since theorem 1 can be verified. It is easier to directly verify that, in all our examples, the m/n bootstrap distribution of $T_n(\hat{F}_n, F)$ converges weakly (in probability) to its limit $\mathcal{L}(F)$ and conclude that theorem 2 holds for all h continuous and bounded than to check the conditions of theorem 2. Such verifications can be found in the papers we cite. We sketch in what follows how the conditions of theorem 1 and 2 can be applied.

Example 1:

a) We sketch heuristically how one would argue for functionals considered in section 2 rather than quantiles. For J_m we need only check that (2.6) holds since $\sqrt{m}(\bar{X} - \mu(F)) = o_p(1)$. For B_m note that the distribution of $m^{-1/2}(i_1 X_1 + \ldots + i_r X_r)$ differs from that of $m^{-1/2}(X_1 + \ldots + X_m)$ by $O(\sum_{j=1}^{r} \frac{(i_j^2 - 1)}{m})$. If we maximize $\sum_{j=1}^{r}(i_j^2 - 1)$ subject to $\sum_{j=1}^{r} i_j = m$, $i_j \geq 1$ we obtain $\frac{2(m-r)}{m} + \frac{(m-r)^2}{m}$. Thus for suitable h, $\delta_m(x) = 2(1-x) + \frac{1}{\sqrt{m}}(1-x)^2$ and the hypotheses of theorem 2 hold.

b) Notice that,

$$P[\sqrt{n}\frac{(\bar{X} - \mu(F))}{s} \leq t] = P[\sqrt{n}(\bar{X} - \mu(F)) - st \leq 0]$$

and apply the previous arguments to $T_n(\hat{F}_n, F) \equiv \sqrt{n}(\bar{X} - \mu(F)) - st$.

Example 2: In example 2 the variance corresponds to $h(x) = x^2$ if $T_m(\hat{F}_m, F) = m^{1/2}(\text{med}(X_1, \ldots, X_m) - F^{-1}(\frac{1}{2}))$. An argument parallel to that in Efron (1979) works. Here is a direct argument for h bounded.

$$(a) \qquad P[\text{med}(X_1^{(i_1)}, \ldots, X_r^{(i_r)}) \neq \text{med}(X_1^{(i_1)}, \ldots X_r^{(i_r - 1)}, X_{r+1})] \leq \frac{1}{r+1}.$$

Thus,

$$(b) \qquad P[\text{med}(X_1^{(i_1)}, \ldots, X_r^{(i_r)}) \neq \text{med}(X_1, \ldots, X_m)]$$

$$\leq \sum_{j=r+1}^{m} \frac{1}{j} \leq \log(\frac{m}{r}).$$

Hence for h bounded,

$$\delta_m(x) \leq \|h\|_\infty \log(\frac{1}{x}).$$

and we can apply theorem 2.

Example 3: Follows by checking (3.2) in theorem 1 and that theorem 2 applies for J_m by arguing as above for B_m. Alternatively, argue as in Athreya and Fukushi (1994). □

Arguments similar to those given so far can be applied to the other examples.

Acknowledgement We are grateful to Jiming Jiang and Adele Cutler for essential programming, to John Rice for editorial comments, and to Kjell Doksum for the Blom reference.

References

Athreya, K.B. (1987). Bootstrap of the mean in the infinite variance case. *Ann. Statist.* **15**, 724-731.

Athreya, K.B. and Fukuchi, J. (1994). Bootstrapping extremes of I.I.D. random variables. Proceedings of Conference on Extreme Value Theory (NIST)

Babu, G.J. and Singh, K. (1985). Edgeworth expansions for sampling without replacement from finite populations. *Journ. Mult. Analysis* **17**, 261-278.

Bentkus, V., Götze, F. and van Zwet, W.R. (1994). An Edgeworth expansion for symmetric statistics. *Tech Report* Univ. of Bielefeld.

Beran, R. (1982). Estimated sampling distributions: The bootstrap and competitors. *Ann. Stat.* **10**, 212-225.

Beran, R. and Srivastava, M.S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *Ann. Statist.* **13**, 95-115.

Bertail, P. (1994). Second order properties of an extrapolated bootstrap without replacement. Submitted to *Bernoulli*.

Bhattacharya, R. and Ghosh, J.K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6**, 434-451.

Bickel, P.J. and Ren, J.J. (1995). The m out of n bootstrap and goodness of fit tests with doubly censored data. Robust Statistics, Data Analysis and Computer Intensive Methods Ed. H. Rieder Lecture Notes in Statistics, Springer Verlag.

Bickel, P.J., Klaassen, C.K., Ritov, Y. and Wellner, J. (1993). Efficient and Adaptive Estimation in Semiparametric Models. Johns Hopkins University Press, Baltimore.

Bickel, P.J. and Yahav, Y. (1988). Richardson extrapolation and the bootstrap. *J. Amer. Statist. Assoc.* **83**, 387-393.

Blom, G. (1976). Some properties of incomplete U statistics. *Biometrika* **63**, 573-580.

Bretagnolle, J. (1983). Lois limites du bootstrap de certaines fonctionelles. *Ann. Inst. H. Poincaré, Ser. B* **19**, 281-296.

David, H.A. (1983). Order Statistics. 2nd edition, J. Wiley, New York.

Deheuvels, P., Mason, D. and Shorack, G. (1993). Some results on the influence of extremes on the bootstrap. *Ann. Inst. H. Poincaré* **29**, 83-103.

DeCiccio, T.J. and Romano, J.P. (1989). The automatic percentile method: Accurate confidence limits in parametric models, *Can. J. Statist.* **17**, 155-169.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields* **95**, 125-140.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1-26.

Efron, B. and Tibshirani, R.J. (1993). An Introduction to the Bootstrap. Chapman & Hall, London, New York.

Feller W. (1968). Probability Theory v1. J. Wiley, New York.

Giné, E. and Zinn, J. (1989). Necessary conditions for the bootstrap of the mean. *Ann. Statist.* **17**, 684-691.

Götze, F. (1993). *Bulletin* I.M.S.

Hall, P. (1992). The Bootstrap and Edgeworth Expansion. Springer Verlag, New York.

Hall, P. and Wilson, S. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* **47**, 757-762.

Hall, P. and Jing B.Y. (1993). Performance of boostrap for heavy tailed distributions. *Tech Report* A.N.U. Canberra.

Hinkley, D.V. (1988). Bootstrap methods (with discussion). *J. Royal. Statist. Soc. B* **50**, 321-337.

Mammen, E. (1992). When does bootstrap work? Springer Verlag, New York.

Politis, D.N. and Romano, J.P. (1994). A general theory for large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22**, 2031-2050.

Praestgaard, J. and Wellner, J. (1993). Exchangeably weighted bootstrap of the general empirical process. *Ann. Probab.* **21**, 2053-2086.

Putter, H. and van Zwet, W.R. (1993). Consistency of plug in estimators with applications to the bootstrap. Submitted to *Ann. Statist.*

Robinson, J. (1978). An asymptotic expansion for samples from a finite population. *Ann. Statist.* **6**, 1005-1011.

Shen, X. and Wong, W. (1994). Convergence rates of sieve estimates. *Ann. Statist.* **22**, 580-615.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9**, 1187-1195.

Wu, C.F.J. (1990). On the asymptotic properties of the jackknife histogram. *Ann. Statist.* **18**, 1438-1452.